Interaktionen und Modifikation von RNA und Proteinen

Prof. Sonja Prohaska

Computational EvoDevo
Universität Leipzig

SS 2018

DNA is “never” naked in a cell

general binder
DNA is usually in association with proteins. In all domains of life there are small, basic **chromatin associated proteins (ChAP)** attached to the DNA forming higher order structures. Their function is to prevent DNA from agglutination, ensuring stability and flexibility, aid structure formation (“packaging”) and engage in gene regulation.

specific binder
Sequence-specific **transcription factors** associate with specific binding sites. There functions is commonly gene regulation.

Proteins that are general binders of DNA

eukaryots

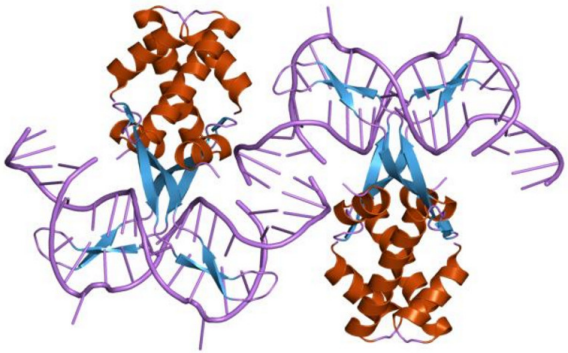
- the DNA-protein complex: chromatin
- major chromatin associating proteins (ChAP): **histones**

bacteria

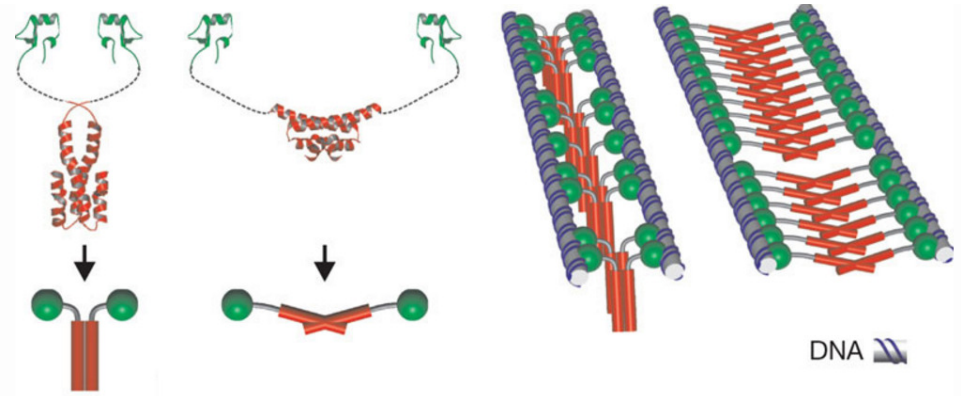
- the DNA-protein complex: nucleoid
- nucleoid associating proteins (NAP):

  HU – introduces negative supercoiling

  H-NS – regulates gene expression



HU, core (red), arms (blue)



histone-like **n**ucleoid-**s**tructuring protein (H-NS)

general vs. specific binder

general binder

- small basic proteins
- contact negatively charged DNA backbone
- e.g. HU, Alba, histones (double stranded)
- e.g. SSB (single stranded)
- bind everywhere, often bend the DNA

specific binder

- large protein with DNA-binding domain(s) (DBD)
- contact major groove of DNA
- DBDs: e.g. HTH (helix-turn-helix), zinc finger, leucine zipper, helix-loop-helix
- bind to specific sites
- sequence-dependent

DNA-binding proteins – a molecular view

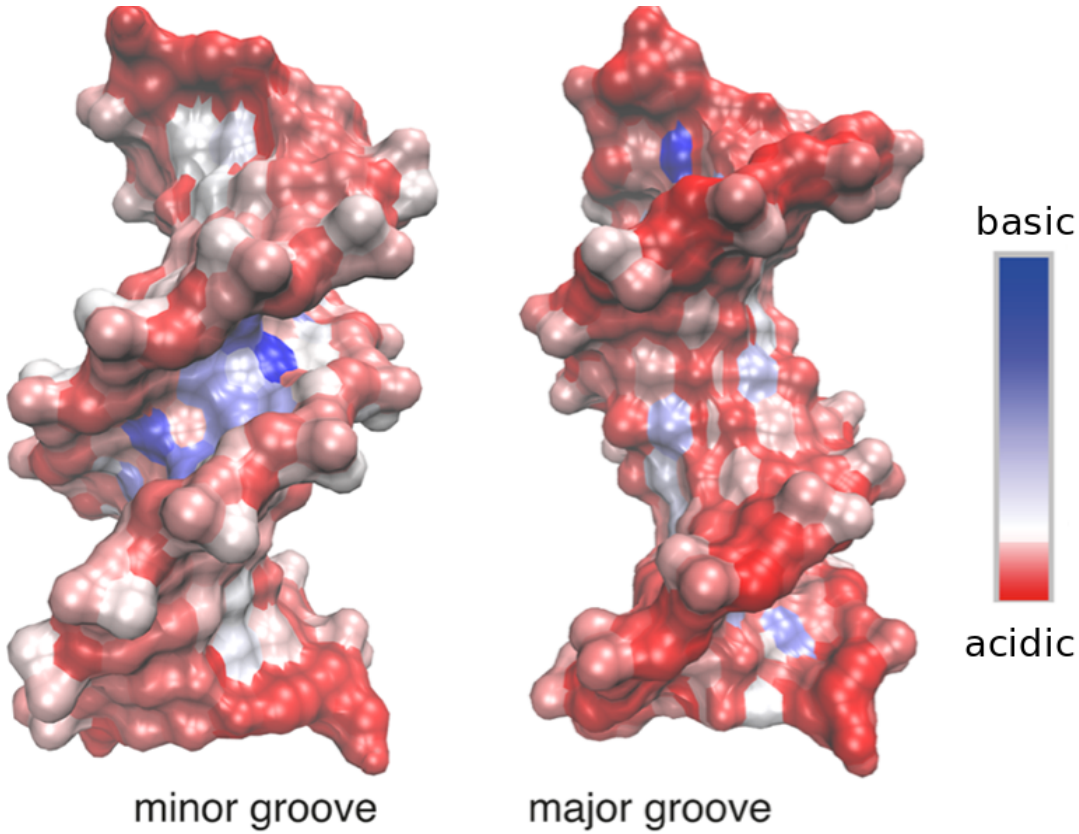
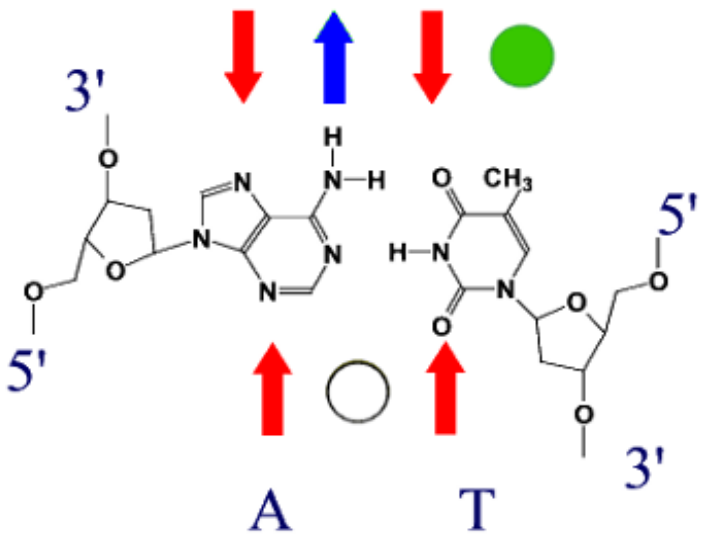
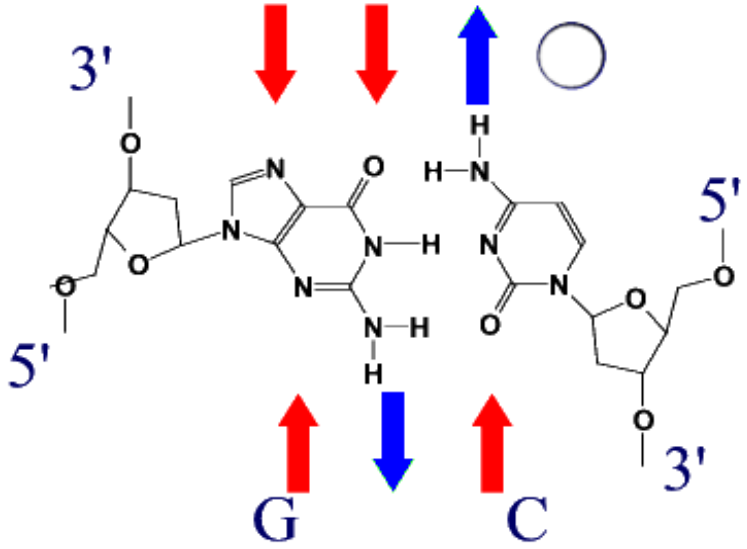


Fig: Electrostatic potential surface of the DNA. General binders are basic and contact the acidic DNA backbone. Specific binders make contact with the nucleobases in the major and minor groove.

Major Groove

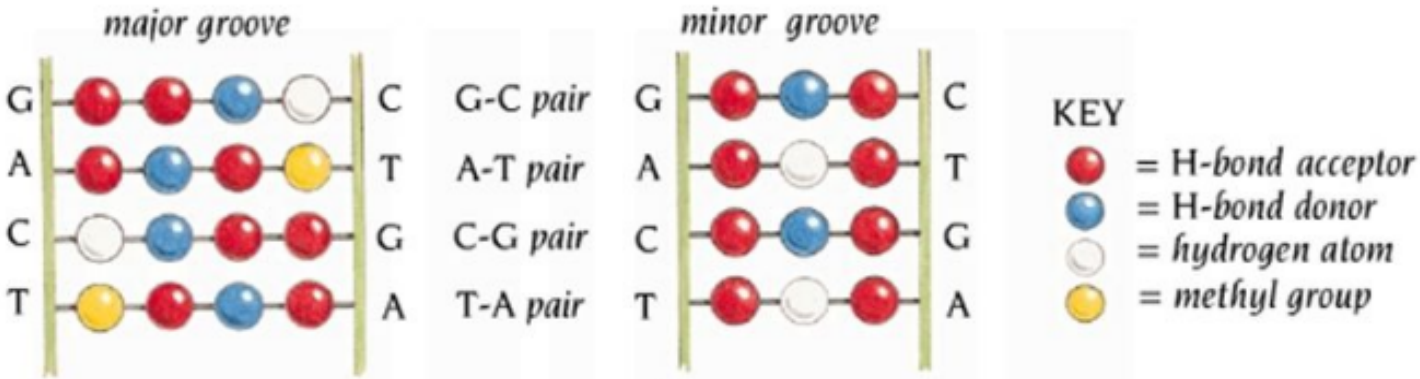


Major Groove



Minor Groove

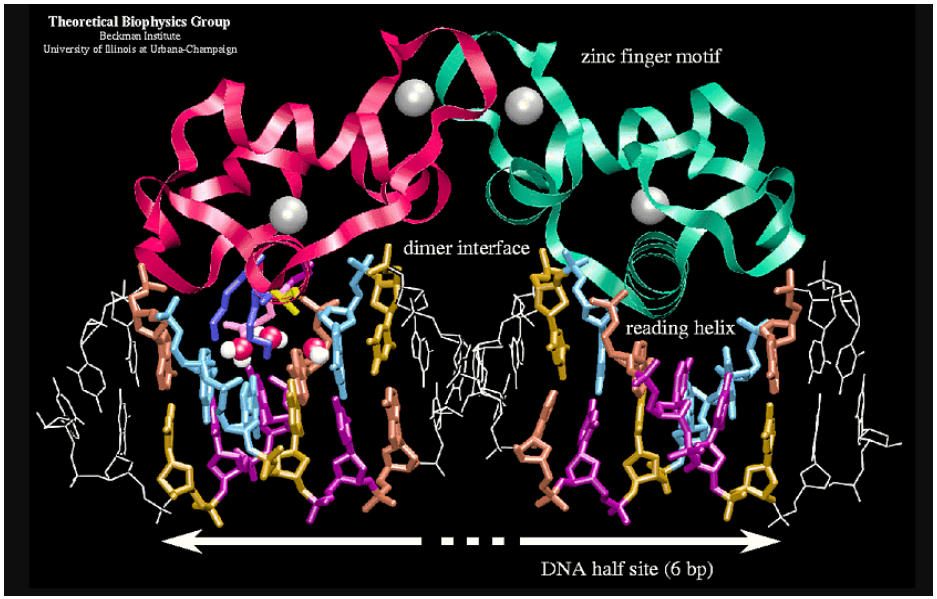
Minor Groove

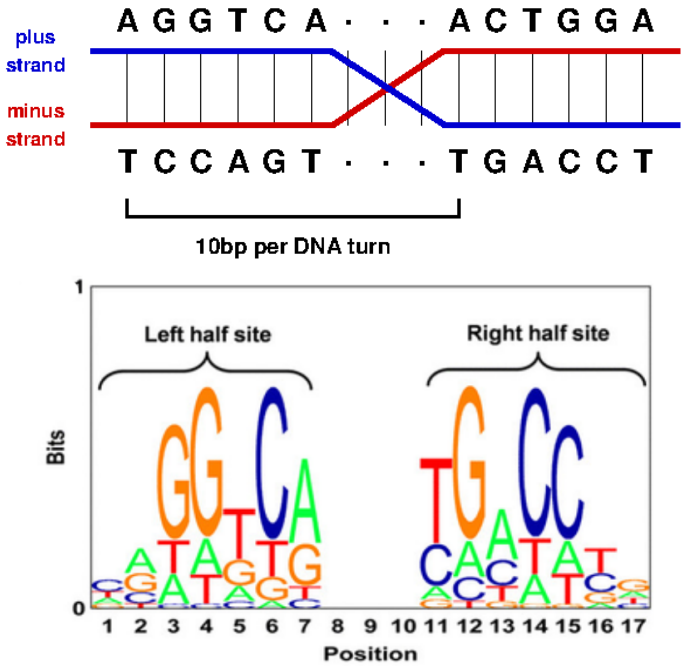


Proteins can distinguish all four nucleobases in the major groove. In the minor groove only AT and GC basepairs can be distinguished.

DNA-binding proteins – a molecular view

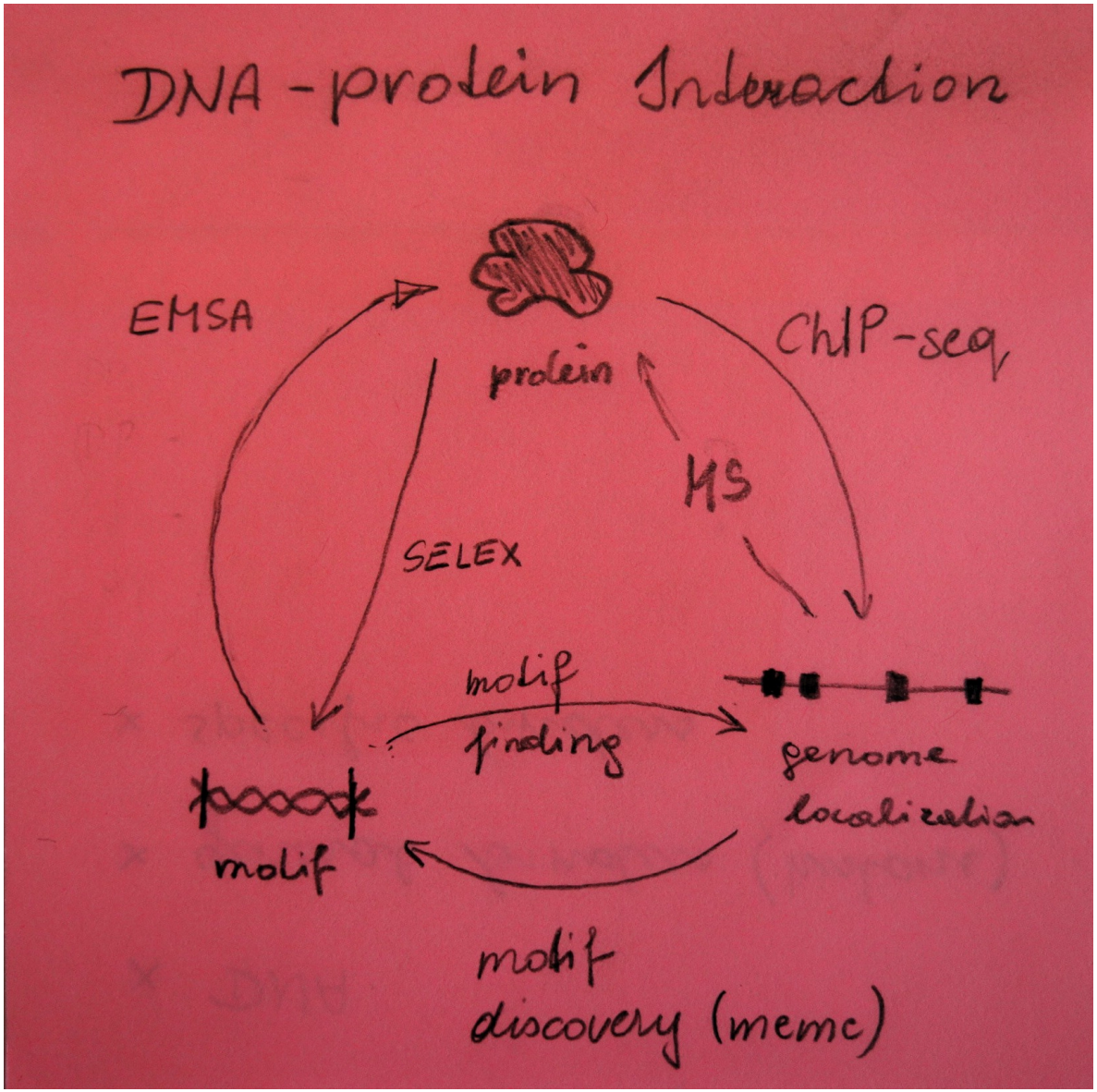
Examle: the estrogen receptor dimer





Homodimeric proteins consist of two identical units that are in contact via the dimer interface. Here each dimer makes contact to 6 base pairs in the major groove in a spequence-specific manner. The motif is **palindromic**.

How to study DNA-protein binding?



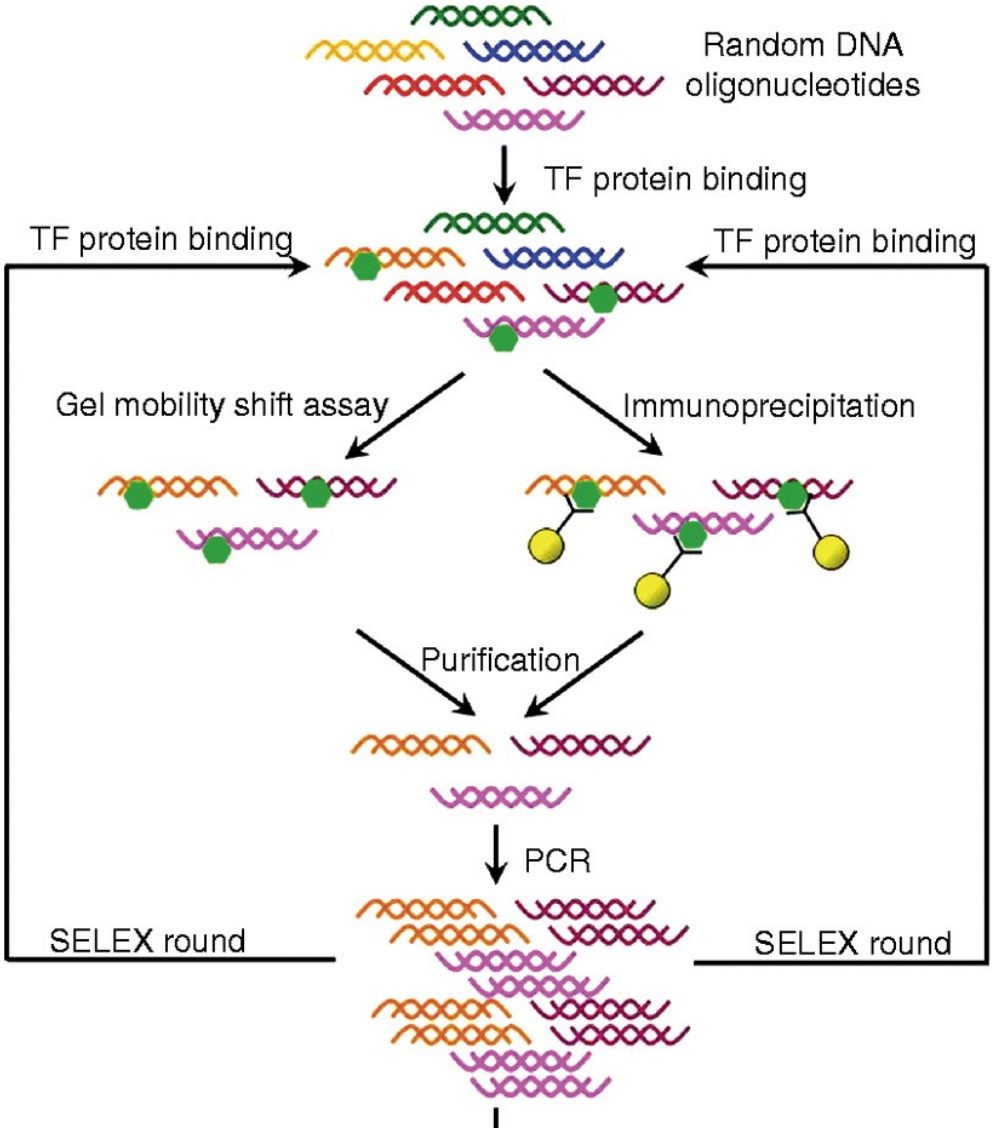
How to study DNA-protein binding?

Given a protein, which DNA sequence does it bind preferentially?

in vitro selection – SELEX

- SELEX (**s**ystematic **e**volution of **l**igands by **ex**ponential enrichment) also known as in-vitro evolution
- a very large oligonucleotide library
  - randomly generated sequences of fixed length
- is exposed to the target protein
- unbound oligos flow through (are removed)
- bound sequences are eluted and amplified by PCR
  - another SELEX round with increased stringency is done or
- the pool of sequences is finally sequenced
- an alignment reveals the **motif** and its variation

How to study DNA-protein binding?



	BamHI		PstI
	---GGATCC...		...CCTGCAGG---
1,	----------G	TGAGTCA	CCATCCCGGTTGGC---
2,	--------GACAGGTC	TGAGTCA	TATGCACG---------
3,	----------GAGCAA	TGAGTCA	TCGTGTCGCG-------
4,	----------GA	TGAGTCA	CAGGCAACGGGCAC---
5,	----------A	TGACTCA	TTGAGCGACTTACCG--
7,	CTGTATTACGACACAA	TGAGTCA	----------
8,	--ATAACGAGTGGGGA	TGACTCA	TT----------
9,	-ATTCTCGCCTTTCTG	TGAGTCA	T----------
11,	----------AAAAAA	TGAGTCA	TCCGAGCT---------
12,	----------AT	TGACTCA	TACGCAGCTAGACT---
13,	-------GAGTGTAGA	TGACTCA	TGGACTG----------
14,	----------A	TGAGTCA	TCGCTAGTCCATGGG--
15,	---GTGTCCTTCGGGA	TGAGTCA	TGC----------
16,	-----GTCACGGGGCC	TGACTCA	TAGAA----------
18,	----------G	TGAGTCA	CGGAAATTGTTGG----
19,	-----GCATTATGGAG	TGACTCA	TCCTT----------
20,	------GAAGCATTTG	TGAGTCA	TCGCT----------
21,	---TCGGTAGTCGGTA	TGAGTCA	TT----------
22,	----------CGG	TGACTCA	CGTAGAGGTAACC----
25,	----------A	TTAGTCA	TCAGAGGGTGGCGAC--
26,	--AAATTTACATGCGA	TGAGTCA	TA----------
27,	--AGCCGCGGTGAGAA	TGAGTCA	TA----------
28,	--TGGCCCCGGGTCTC	TGACTCA	GC----------
29,	----AACGAGACGCGC	TGAGTCA	TCTT----------
30,	--GCCAGTTGATACTC	TGTGTCA	CGG----------
32,	----GTAGCTGAGGAG	TGAGTCA	CGCCTG----------
33,	----------GGGG	TGAGTCA	TAAAGATAAATCT----
34,	-CCCGCGTAGGCTCGA	TGAGTCA	A----------
35,	--TCAGAGTCAGCTTA	TGAGTCA	GG----------
36,	----------CGCTGG	TGACTCA	TCGTGTTCTG-------
37,	--GCAGGCGCCACGCG	TGACTCA	T----------
38,	-------GTCAGTTTG	TGAGTCA	CTCTACC----------
39,	----GTATGACGTGGA	TGTGTCA	GAGG----------
40,	--------TGGCCCTA	TGACTCA	TAAGGCAC---------
41,	--GCCAATGACTTCTC	TGAGTCA	T----------
42,	--CCGTCATCGCGGGT	TGAGTCA	CT----------
43,	----------GA	TGAGTCA	GGACCGGGGTTGGA---
44,	----------TACGTG	TGACTCA	TTCACCACCC-------
45,	----------GCCTG	TGACTCA	TCCTGCGCGTA------
46,	----------GAAATA	TGAGTCA	CGGGACGTCT-------
47,	-AGTGTGTAAAGGGGG	TGAGTCA	T----------
48,	----------AG	TGACTCA	CCGACACGTACGT----
49,	-------TGCGTCGGG	TGAGTCA	GATTGTG----------

FIG. 3. GCN4-binding sites.

How to represent a motif?

Given an alignment of n motifs of length l
here the result of the SELEX experiment: $n = 43$, $l = 7$

- **consensus string**

  most frequent nucleotide per position 5'-TGAGTCA-3'

- **consensus pattern**

  all possible nucleotides per position 5'-T(G|T)(A|T)(G|C)TCA-3'

  $K = \{G, T\}$, $W = \{A, T\}$, $S = \{G, C\} \rightarrow$ 5'-TKWSTCA-3'

- **position frequency matrix – PFM**

  Alphabeth $= \{A, C, G, T\}$, size of Alphabet a, matrix $a \times l$

- **sequence logo** bast on the information content (IC)

How to calculate the information content?

The information content (y-axis) is measured in bits.
The maximal information depends on the size k of the Alphabet $\mathfrak{A}$,
here $\mathfrak{A} = \{A, C, G, T\}$, $a \in \mathfrak{A}$, and $k = 4$.
The **Shannon entropy** (uncertainty) H_i at position i of the motif is

$$H_i = -\sum_{\mathfrak{A}} f_i(a) log_2 f_i(a) \tag{1}$$

where $f_i(a)$ is the relative frequency of nucleotide a at position i.
The **information content** R_i at position i of the motif is

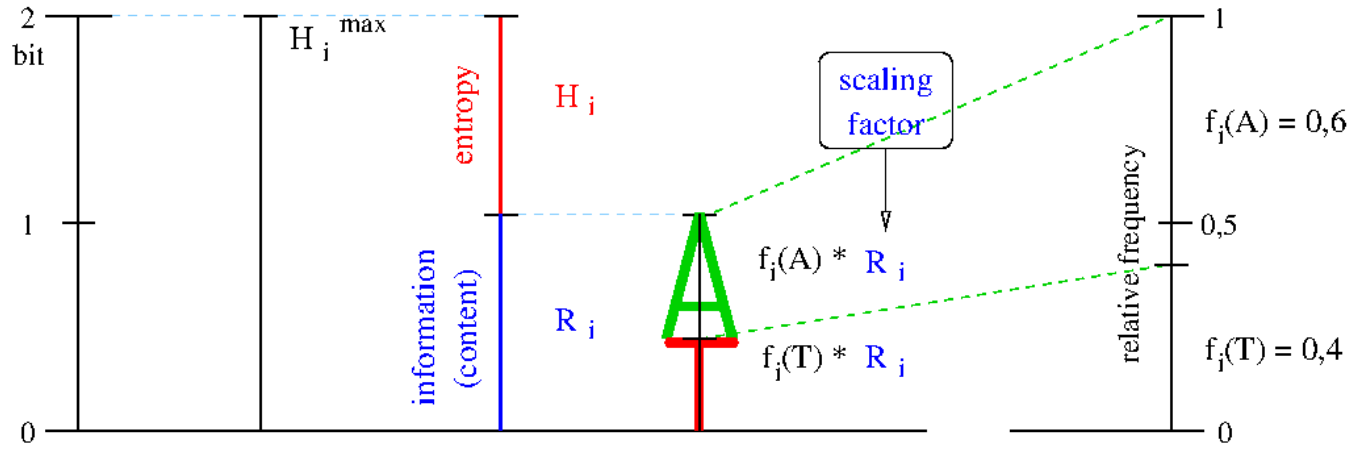
$$R_i = log_2(k) - H_i \tag{2}$$

(i.e. the maximal information content minus the uncertainty)

The height $h_i(a)$ of a letter a in column i is

$$h_i = f_i(a) \times R_i \tag{3}$$

note: $log_2 k = log_2 4 = 2$



... with small-sample size correction

small-sample correction e_n is given by

$$e_n = \frac{1}{\ln 2} \times \frac{s-1}{2n} \tag{4}$$

where s is the number of sequences in the alignment.
The information content R_i at position i of the motif is then

$$R_i = log_2(k) - (H_i + e_n) \tag{5}$$

Explain, what does e_n do?

What to do with the motif?

How often would you expect to find a motif m in a sequence M?

- assume that motif m is a string, e.g. $5'$-GGCCT-$3'$
- of motif length $l = 5$bp
- the sequence M, e.g. the human HoxA cluster,
- has a length of $L = 163001$bp
- Alphabet $\mathfrak{A} = \{A, C, G, T\}$

1. assume **uniform** nucleotide distribution: $f(x) = 0.25$
2. use **mono**-nucleotide distribution:
   $f(A) = 0.2428, f(C) = 0.2555, f(G) = 0.2552, f(T) = 0.2466$
3. use **di**-nucleotide distribution:
   $f(AG) = 0.0604, f(GG) = 0.0812, f(GC) = 0.0677$
   $f(CC) = 0.0815, f(CT) = 0.0751$

Calculations

$$E^{UNI}(m) = f(X)^5 \times (L - (l - 1)) \tag{6}$$

$$E^{MONO}(m) = (f(G)^2 \times f(C)^2) \times f(T)) \times (L - (l - 1)) \tag{7}$$

$$E^{DI}(m) = \frac{f(GG) \times f(GC) \times f(CC) \times f(CT)}{f(G) \times f(C)^2} \times (L - (l - 1)) \tag{8}$$

Results: $E^{UNI}(m) = 159$; $E^{MONO}(m) = 171$; $E^{DI}(m) = 329$;

Which expectation comes closer to the observation?

Let's assume you apply a word search method to the human hox cluster sequence and find 312 motif sites.

Observed sites: number of matches of motif m in M is $O = 312$

Expected sites: $E^{UNI}(m) = 159; E^{MONO}(m) = 171; E^{DI}(m) = 329;$

Use the **chi-squared test (χ^2)** to determine whether there is a significant difference between the expected frequency and the observed frequency.

$$\chi^2 = \frac{(O - E)^2}{E}$$

df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750

Results: $\chi^2(UNI) = 147.226; \chi^2(MONO) = 116.263; \chi^2(DI) = 0.878;$

Which expectation comes closer to the observation?

Significance levels

for **one** degree of freedom (df)

expected and observed value, E and O, are significantly different

at 5% level ($p = 0.05$) if $\chi^2 >= 3.841$

at 1% level ($p = 0.01$) if $\chi^2 >= 6.635$

Conclusion

$E^{UNI}(m)$ and $E^{MONO}(m)$ are significantly different from O,

$E^{DI}(m) = 329$ is not.

Therefore we can conclude that the calculation of $E(m)$ using the di-nucleotide distribution is a good prediction for the **observable** number of sites.

DNA is more than a (plus) strand!

- DNA is double stranded
- the two strands are refered to as **plus and minus strand**
- the **convention** is:
- the strand which runs from 5' to 3' from left to right
- is the top strand (when both strands are displayed) and
- it is also the plus strand (the one displayed if only one strand is displayed)
- the sequence of one strand determines the sequence of the other
- the relation between the two strands is the "reverse complement"

about motifs and sites

- the (binding) motif specifies a (binding) pattern
- a sequence in the genome matching the motif is a (binding) site

Exercises

- Explain why the expected number of sites E comes closer to the true number of sites O in the sequence M when using the dinucleotide instead of mononucleotide composition.
- How would you need to adapt equations 6 to 8 to account for both DNA strands in the calculation of the expecation value? What if the motif is palindromic, e.g. 'GGTACC'?
- How many different motifs of lenght 5 can be derived from the sequence below?
- How many binding sites can be found with motif 'GGTCA'